模块四 成对数据的统计分析

第 1 节 一元线性回归模型及其应用（★★★）

强化训练

1.（2023 • 河南模拟 • ★）为了研究汽车减重对降低油耗的作用，对一组样本数据$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$进行分析，其中$x_i$表示减重质量（单位：kg），$y_i$表示每行驶一百公里降低的油耗（单位：升），$i=1,2,\cdots,n$，由此得到的经验回归方程为$\hat{y}=\hat{b}x+\hat{a}(\hat{b}>0)$，有下列四个说法：

①$\hat{a}$的值一定为 0；②$\hat{b}$越大，减重对降低油耗的作用越大；③决定系数R^2越大，拟合效果越好；④至少有一个数据点在经验回归直线上. 其中所有正确说法的编号是（　　）

（A）①④　　（B）②③　　（C）②③④　　（D）①②④

答案：B

解析：①项，从实际意义来看，a表示减重 0kg 时，每行驶一百公里降低的油耗，应该为 0，但$\hat{a}$是由样本观测数据求出的a的估计值，只会近似等于 0，但不是一定为 0，故①项错误；

②项，$\hat{b}$越大，则每减重 1kg，行驶一百公里降低的油耗越多，减重对降低油耗的作用越大，故②项正确；

③项，决定系数R^2越大，则残差平方和越小，拟合效果越好，故③项正确；

④项，用最小二乘法估计经验回归方程，数据点不一定落在经验回归直线上，故④项错误.

2.（2023 • 湖南模拟 • ★）（多选）设某大学的女生体重y（单位：kg）与身高x（单位：cm）具有线性相关关系，根据一组样本数据$(x_i,y_i)(i=1,2,\cdots,n)$，用最小二乘法建立的经验回归方程为$\hat{y}=0.85x-85.71$，则下列结论中正确的是（　　）

（A）y与x有正的线性相关关系

（B）若该大学女生的平均身高为 168cm，则平均体重约为 57.09kg

（C）若该大学某女生身高增加 1cm，则其体重约增加 0.85kg

（D）若该大学某女生身高为 170cm，则可断定其体重必为 58.79kg

答案：ABC

解析：A 项，经验回归直线的斜率$\hat{b}=0.85>0$，所以y与x正相关，故 A 项正确；

B 项，当$x=168$时，$\hat{y}=0.85\times168-85.71=57.09$，所以平均体重约为 57.09kg，故 B 项正确；

C 项，因为$\hat{b}=0.85$，所以当女生身高增加 1cm 时，其体重约增加 0.85kg，故 C 项正确；

D 项，看到题干的“其体重必为 58.79kg”，即可得出 D 选项错误，因为只能说“其体重约为 58.79kg”.

3.（2023 • 宁夏吴忠模拟 • ★★）某种产品的广告费用x（单位：万元）与销售额y（单位：万元）之间的关系如下表，若y与x的经验回归方程为$\hat{y}=1.3x+m$，则$m=$（　　）

x	1	3	4	5	7
y	6	8	12	10	14

（A）4.1　　（B）4.7　　（C）4.8　　（D）6.8

答案：C

解析： 经验回归方程中只有 m 未知，可将样本中心点 $(\bar{x},\bar{y})$ 代入，

由题意，$\bar{x}=\dfrac{1+3+4+5+7}{5}=4$，$\bar{y}=\dfrac{6+8+12+10+14}{5}=10$，

将 $(4,10)$ 代入 $\hat{y}=1.3x+m$ 可得 $10=1.3\times4+m$，所以 $m=4.8$.

4.（2023·吉林模拟·★★）某地以“绿水青山就是金山银山”理念为引导，推进绿色发展，现要订购一批苗木，苗木长度与售价如下表：

苗木长度 x（cm）	38	48	58	68	78	88
售价 y（元）	16.8	18.8	20.8	22.8	24	25.8

若苗木长度 x（cm）与售价 y（元）之间存在线性相关关系，其经验回归方程为 $\hat{y}=\hat{b}x+8.9$，则当售价大约为 38.9 元时，苗木长度大约为（　　）

（A）148cm　　（B）150cm　　（C）152cm　　（D）154cm

答案： B

解析： 应先由表中数据求出 $\hat{b}$，才能作出估计，可将样本中心点 $(\bar{x},\bar{y})$ 代入经验回归方程，

由题意，$\bar{x}=\dfrac{38+48+58+68+78+88}{6}=63$，$\bar{y}=\dfrac{16.8+18.8+20.8+22.8+24+25.8}{6}=21.5$，

将 $(63,21.5)$ 代入 $\hat{y}=\hat{b}x+8.9$ 可得 $21.5=63\hat{b}+8.9$，解得：$\hat{b}=0.2$，所以 $\hat{y}=0.2x+8.9$，

当 $\hat{y}=38.9$ 时，$38.9=0.2x+8.9$，解得：$x=150$，所以当售价大约为 38.9 元时，苗木长度大约为 150cm.

5.（2022·贵州模拟·★★★）某企业新研发了一种产品，产品的成本由原料成本及非原料成本组成，每件产品的非原料成本 y（元）与生产的产品数量 x（千件）有关，经统计得到如下数据：

x	2	5	8	9	11
y	12	10	8	8	7

（1）根据表中的数据，运用相关系数进行分析说明，可以用一元线性回归模型拟合 y 与 x 的关系，并指出是正相关还是负相关；

（2）求 y 关于 x 的经验回归方程，并预测生产该产品 13 千件时，每件产品的非原料成本为多少元？

参考公式：相关系数 $r=\dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}}$，经验回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的 $\hat{b}$ 和 $\hat{a}$ 的最小二乘估计公式为 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}$，$\hat{a}=\bar{y}-\hat{b}\bar{x}$；参考数据：$\sqrt{2}\approx1.414$.

解：（1）（直接代给出的相关系数公式求 r 可行，但注意到数据的绝对值都不大，按 $r=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{(\sum\limits_{i=1}^{n}x_i^2-n\bar{x}^2)\cdot(\sum\limits_{i=1}^{n}y_i^2-n\bar{y}^2)}}$ 来算更方便，先求 $\bar{x}$ 和 $\bar{y}$）

由表中数据可得，$\overline{x}=\dfrac{2+5+8+9+11}{5}=7$，$\overline{y}=\dfrac{12+10+8+8+7}{5}=9$，

所以$\sum\limits_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})=\sum\limits_{i=1}^{5}x_iy_i-5\overline{x}\,\overline{y}=2\times12+5\times10+8\times8+9\times8+11\times7-5\times7\times9=-28$，

$\sum\limits_{i=1}^{5}(x_i-\overline{x})^2=\sum\limits_{i=1}^{5}x_i^2-5\overline{x}^2=2^2+5^2+8^2+9^2+11^2-5\times7^2=50$，

$\sum\limits_{i=1}^{5}(y_i-\overline{y})^2=\sum\limits_{i=1}^{5}y_i^2-5\overline{y}^2=12^2+10^2+8^2+8^2+7^2-5\times9^2=16$，

故相关系数$r=\dfrac{\sum\limits_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{5}(x_i-\overline{x})^2}\cdot\sqrt{\sum\limits_{i=1}^{5}(y_i-\overline{y})^2}}=\dfrac{-28}{\sqrt{50}\times\sqrt{16}}=-\dfrac{7}{5\sqrt{2}}\approx-\dfrac{7}{5\times1.414}\approx-0.99$，

因为$|r|$很接近 1，所以x，y的线性相关性很强，可用一元线性回归模型拟合y与x的关系，

又$r<0$，所以y与x负相关.

（2）由（1）可得$\hat{b}=\dfrac{\sum\limits_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{5}(x_i-\overline{x})^2}=\dfrac{-28}{50}=-0.56$，$\hat{a}=\overline{y}-\hat{b}\overline{x}=9-(-0.56)\times7=12.92$，

所以y关于x的经验回归方程为$\hat{y}=-0.56x+12.92$，当$x=13$时，$\hat{y}=-0.56\times13+12.92=5.64$，

故可预测生产该产品 13 千件时，每件产品的非原料成本约为 5.64 元.

【反思】若观测数据x_i，y_i的绝对值较小，则按$\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\overline{y}$，$\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2$，$\sum\limits_{i=1}^{n}y_i^2-n\overline{y}^2$来算比按$\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$，$\sum\limits_{i=1}^{n}(x_i-\overline{x})^2$，$\sum\limits_{i=1}^{n}(y_i-\overline{y})^2$计算简单，否则宜采用后者计算.

6．（2023•江苏泰州模拟•★★★）小家电指除大功率、大体积家用电器（如冰箱、洗衣机、空调等）以外的家用电器，运用场景广泛，近年来随着科技发展，智能小家电市场规模呈持续发展趋势，下表为连续 5 年中国智能小家电市场规模（单位：千亿元），其中年份对应的代码依次为 1~5.

年份代码x	1	2	3	4	5
市场规模y	0.9	1.2	1.5	1.4	1.6

（1）由上表数据可知，可用线性回归模型拟合y与x的关系，请用相关系数加以说明；

（2）建立y关于x的经验回归方程.

参考数据：$\overline{y}=1.32$，$\sum\limits_{i=1}^{5}x_iy_i=21.4$，$\sqrt{\sum\limits_{i=1}^{5}(y_i-\overline{y})^2}\approx0.55$，$\sqrt{10}\approx3.16$；

参考公式：相关系数$r=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}}$，经验回归方程$\hat{y}=\hat{b}x+\hat{a}$中的$\hat{b}$和$\hat{a}$的最小二乘估计

公式分别为$\hat{b}=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$.

解：（1）（观察所给公式发现求相关系数 r，还差$\sum_{i=1}^{5}(x_i-\overline{x})^2$和$\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})$，故分别计算）

由题意，$\overline{x}=\dfrac{1+2+3+4+5}{5}=3$，所以$\sum_{i=1}^{5}(x_i-\overline{x})^2=(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2=10$，

（再算$\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})$，参考数据中有$\sum_{i=1}^{5}x_iy_i$，故转化为$\sum_{i=1}^{5}x_iy_i-5\overline{x}\,\overline{y}$来算）

$\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{5}x_iy_i-5\overline{x}\,\overline{y}=21.4-5\times3\times1.32=1.6$，

所以$r=\dfrac{\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{5}(x_i-\overline{x})^2}\cdot\sqrt{\sum_{i=1}^{5}(y_i-\overline{y})^2}}\approx\dfrac{1.6}{\sqrt{10}\times0.55}\approx\dfrac{1.6}{3.16\times0.55}\approx0.92$，

故 y 与 x 的线性相关程度较高，可用线性回归模型拟合 y 与 x 的关系.

（2）（算$\hat{b}$需要用到的数据都有了，故直接计算）由（1）可得$\hat{b}=\dfrac{\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{5}(x_i-\overline{x})^2}=\dfrac{1.6}{10}=0.16$，

$\hat{a}=\overline{y}-\hat{b}\overline{x}=1.32-0.16\times3=0.84$，所以 y 关于 x 的经验回归方程为$\hat{y}=0.16x+0.84$.

7.（2023•江苏苏州模拟•★★★）新能源汽车作为战略新兴产业，代表汽车产业的发展方向，发展新能源汽车，对改善能源消费结构、减少空气污染、推动汽车产业和交通运输业转型升级具有积极意义，经过十多年的精心培育，我国新能源汽车产业取得了显著成绩，产销量连续四年全球第一，保有量居全球首位.

（1）已知某公司生产的新能源汽车电池的使用寿命ξ（单位：万公里）服从正态分布$N(60,16)$，问：该公司每月生产的 2 万块电池中，大约有多少块电池的使用寿命可以超过 68 万公里？

参考数据：若随机变量$\xi\sim N(\mu,\sigma^2)$，则$P(\mu-\sigma<\xi<\mu+\sigma)\approx0.683$，$P(\mu-2\sigma<\xi<\mu+2\sigma)\approx0.955$，$P(\mu-3\sigma<\xi<\mu+3\sigma)\approx0.997$.

（2）下表给出了我国 2017~2021 年新能源汽车保有量 y（单位：万辆）的数据.

年份	2017	2018	2019	2020	2021
年份代码 x	1	2	3	4	5
新能源汽车保有量 y	153	260	381	492	784

经计算，变量 x，y 的样本相关系数$r_1\approx0.946$，变量x^2与 y 的样本相关系数$r_2\approx0.985$.

①试判断$\hat{y}=\hat{b}x+\hat{a}$和$\hat{y}=\hat{b}x^2+\hat{a}$哪一个更适合作为 y 与 x 之间的回归模型？

②根据①的判断结果，求出 y 关于 x 的回归方程（精确到 0.1），并预测 2023 年我国新能源汽车的保有量.

参考数据：令$t_i=x_i^2(i=1,2,3,4,5)$，计算得$\overline{y}=414$，$\sum_{i=1}^{5}x_iy_i=7704$，$\sum_{i=1}^{5}t_iy_i=32094$，$\sum_{i=1}^{5}t_i^2=979$.

参考公式：在回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中，$\hat{b}=\dfrac{\sum_{i=1}^{n}x_iy_i-n\overline{x}\overline{y}}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$.

解：（1）因为 $\xi\sim N(60,16)$，该正态分布的均值 $\mu=60$，标准差 $\sigma=4$，

所以 $P(\xi>68)=P(\xi>\mu+2\sigma)=\dfrac{1-P(\mu-2\sigma<\xi<\mu+2\sigma)}{2}\approx\dfrac{1-0.955}{2}=0.0225$，

故该公司每月生产的 2 万块电池中，使用寿命可以超过 68 万公里的块数约为 $20000\times0.0225=450$.

（2）①因为 $r_1\approx0.946$，$r_2\approx0.985$，所以 $|r_2|>|r_1|$，故 $\hat{y}=\hat{b}x^2+\hat{a}$ 更适合作为 y 与 x 之间的回归模型.

②令 $t=x^2$，则 $\hat{y}=\hat{b}x^2+\hat{a}$ 即为 $\hat{y}=\hat{b}t+\hat{a}$，（y 关于 t 为线性回归模型，可用最小二乘估计公式求 $\hat{b}$ 和 $\hat{a}$）

由题意，$\overline{t}=\dfrac{t_1+t_2+\cdots+t_5}{5}=\dfrac{x_1^2+x_2^2+\cdots+x_5^2}{5}=\dfrac{1^2+2^2+\cdots+5^2}{5}=11$，

结合参考数据可得 $\hat{b}=\dfrac{\sum_{i=1}^{5}t_iy_i-5\overline{t}\,\overline{y}}{\sum_{i=1}^{5}t_i^2-5\overline{t}^2}=\dfrac{32094-5\times11\times414}{979-5\times11^2}=\dfrac{9324}{374}\approx24.9$，$\hat{a}=\overline{y}-\hat{b}\overline{t}=414-\dfrac{9324}{374}\times11\approx139.8$，

所以 y 关于 t 的回归方程为 $\hat{y}=24.9t+139.8$，故 y 关于 x 的回归方程为 $\hat{y}=24.9x^2+139.8$，

当 $x=7$ 时，$\hat{y}=24.9\times7^2+139.8=1359.9$，所以预测 2023 年我国新能源汽车的保有量约为 1359.9 万辆.

【反思】上述求 $\hat{a}$ 的过程中，若把 $\hat{b}$ 代成近似后的数据 24.9，则求得的 $\hat{a}$ 为 140.1，此时 2023 年我国新能源汽车的保有量的预测值则为 1360.2 万辆，这一结果也算正确.

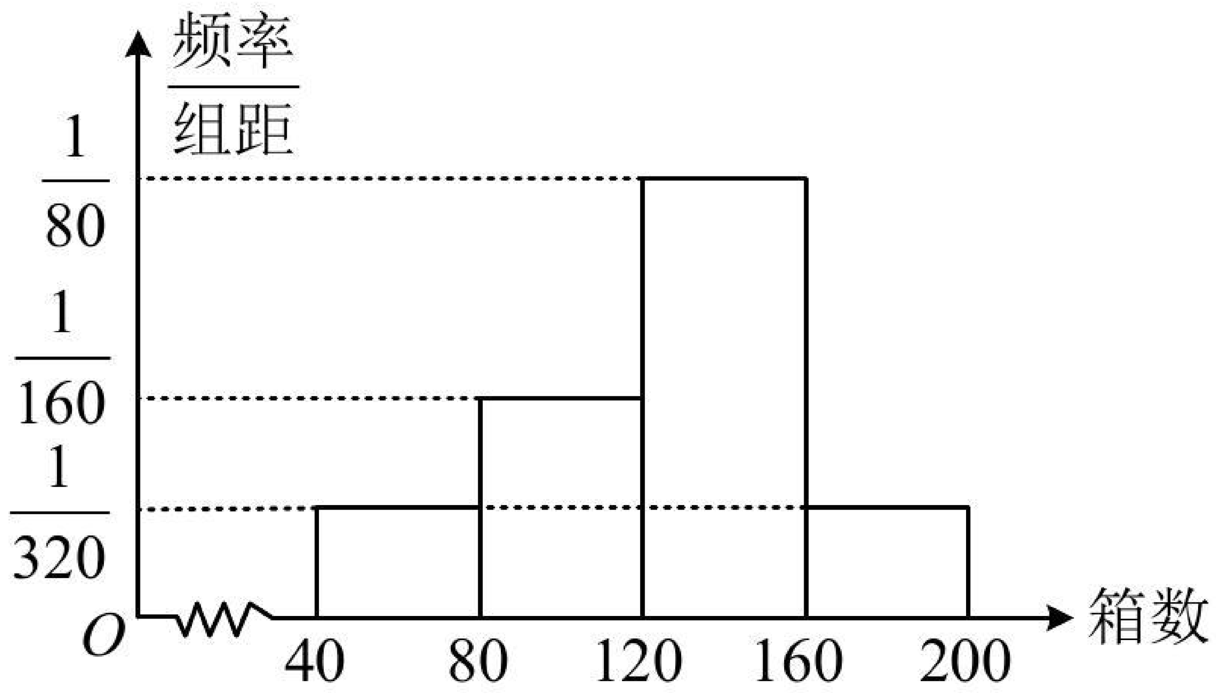
8．（2023•湖南长沙雅礼中学模拟•★★★★）为贯彻中共中央、国务院 2023 年一号文件，某单位在当地定点帮扶某村种植一种草莓，并把这种露天种植的草莓搬到了大棚里，收到了很好的经济效益. 根据资料显示，产出的草莓的箱数 x（单位：箱）与成本 y（单位：千元）的关系如下：

x	1	3	4	6	7
y	5	6.5	7	7.5	8

可用回归方程 $\hat{y}=\hat{b}\lg x+\hat{a}$（其中 $\hat{a}$，$\hat{b}$ 为常数）来拟合 y 与 x 的关系.

（1）若农户卖出该草莓的价格为 150 元/箱，试预测该草莓 100 箱的利润是多少元；（利润 = 售价 − 成本）

（2）据统计，1 月份的连续 16 天中农户每天为甲地可配送的该草莓的箱数的频率分布直方图如图，用这 16 天的情况来估计相应的概率. 一个运输户拟购置 n 辆小货车专门运输农户为甲地配送的该草莓，一辆货车每天只能运一趟，每辆车每趟最多只能装载 40 箱该草莓，满载发车，否则不发车. 若发车，则每辆车每趟可获利 500 元；若未发车，则每辆车每天平均亏损 200 元. 试比较 $n=3$ 和 $n=4$ 时，此项业务每天的利润平均值的大小.



参考数据与公式：线性回归直线 $\hat{y}=\hat{b}t+\hat{a}$ 中，$\hat{b}=\dfrac{\sum_{i=1}^{n}(t_i-\overline{t})(y_i-\overline{y})}{\sum_{i=1}^{n}(t_i-\overline{t})^2}$，$\hat{a}=\overline{y}-\hat{b}\overline{t}$；设 $t=\lg x$，则

$\overline{t}$	$\overline{y}$	$\sum_{i=1}^{5}(t_i-\overline{t})(y_i-\overline{y})$	$\sum_{i=1}^{5}(t_i-\overline{t})^2$
0.54	6.8	1.53	0.45

解：（1）（先求 100 箱草莓的成本，需建立 y 关于 x 的回归方程，题干已经作了变换 $t=\lg x$，故直接求 $\hat{b}$ 和 $\hat{a}$）

由题意，$\hat{b}=\dfrac{\sum_{i=1}^{5}(t_i-\overline{t})(y_i-\overline{y})}{\sum_{i=1}^{5}(t_i-\overline{t})^2}=\dfrac{1.53}{0.45}=3.4$，$\hat{a}=\overline{y}-\hat{b}\overline{t}=6.8-3.4\times0.54=4.964$，

所以 y 关于 t 的回归方程为 $\hat{y}=3.4t+4.964$，又 $t=\lg x$，所以 y 关于 x 的回归方程为 $\hat{y}=3.4\lg x+4.964$，

从而当 $x=100$ 时，$\hat{y}=3.4\lg100+4.964=11.764$，故可预测该草莓 100 箱的利润为 $100\times150-11.764\times1000=3236$ 元.

（2）（利润受每辆车是否发车影响，每辆车是否发车又由可配送的草莓箱数决定，箱数为随机变量，由频率分布直方图可获得其概率分布，从而得到利润的概率分布，下面分别考虑）

设该草莓可配送的箱数为随机变量 X，由频率分布直方图可知 X 的概率分布如下表：

X	$[40,80)$	$[80,120)$	$[120,160)$	$[160,200]$
P	$\dfrac{1}{8}$	$\dfrac{1}{4}$	$\dfrac{1}{2}$	$\dfrac{1}{8}$

当 $n=3$ 时，设此项业务的利润为 Y，则当 $40\le X<80$ 时，只能发 1 辆车，

此时 $Y=500-2\times200=100$，所以 $P(Y=100)=\dfrac{1}{8}$，

当 $80\le X<120$ 时，可发 2 辆车，此时 $Y=500\times2-200=800$，所以 $P(Y=800)=\dfrac{1}{4}$，

当 $120\le X\le200$ 时，可发 3 辆车，此时 $Y=500\times3=1500$，所以 $P(Y=1500)=\dfrac{1}{2}+\dfrac{1}{8}=\dfrac{5}{8}$，

故 $E(Y)=100\times\dfrac{1}{8}+800\times\dfrac{1}{4}+1500\times\dfrac{5}{8}=1150$；

当 $n=4$ 时，设此项业务的利润为 Z，则当 $40\le X<80$ 时，只能发 1 辆车，

此时 $Z=500-3\times200=-100$，所以 $P(Z=-100)=\frac{1}{8}$，

当 $80\le X<120$ 时，可发 2 辆车，此时 $Z=500\times2-200\times2=600$，所以 $P(Z=600)=\frac{1}{4}$，

当 $120\le X<160$ 时，可发 3 辆车，此时 $Z=500\times3-200=1300$，所以 $P(Z=1300)=\frac{1}{2}$，

当 $160\le X\le200$ 时，可发 4 辆车，此时 $Z=500\times4=2000$，所以 $P(Z=2000)=\frac{1}{8}$，

故 $E(Z)=-100\times\frac{1}{8}+600\times\frac{1}{4}+1300\times\frac{1}{2}+2000\times\frac{1}{8}=1037.5$；

因为 $E(Z)<E(Y)$，所以购置 3 辆小货车此项业务的平均利润更大.